Universidade de Vigo

A new flexible direct ROC regression model. Detection of cardiovascular risk factors by anthropometric measures.

María X. Rodríguez-Álvarez, Javier Roca-Pardiñas and Carmen Cadarso-Suárez

Report 09/06

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

A new flexible direct ROC regression model. Detection of cardiovascular risk factors by anthropometric measures.

María X. Rodríguez-Álvarez, Javier Roca-Pardiñas and Carmen Cadarso-Suárez

Report 09/06

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: 

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

A new flexible direct ROC regression model. Detection of cardiovascular risk factors by anthropometric measures.

María X. Rodríguez-Álvarez[1,3], Javier Roca-Pardiñas[2], Carmen Cadarso-Suárez[1,3]

[1] Unit of Biostatistics, Dept. of Statistics and OR. University of Santiago de Compostela.

[2] Dept. of Statistics and OR. University of Vigo.

[3] Instituto de Investigación Sanitaria de Santiago (IDIS). Santiago de Compostela.

February 5, 2010

Abstract

The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the accuracy of diagnostic tests in terms of differentiating between two conditions. It is well known that, in certain circumstances, the characteristics of the patient or the place where the diagnostic test is performed can modify the test's accuracy. This study presents a new estimator for the conditional ROC curve, based on direct modeling. In this approach, the effect of the covariates and false positive fraction on the ROC curve is modeled nonparametrically using generalized additive models (GAM) combined with local polynomial kernel smoothers. Our method allows for incorporation of more than one covariate in the regression model for the ROC curve, and the possible interaction between them. The proposed model's performance is examined in an in-depth simulation study. Lastly, endocrine data are analyzed with the aim of assessing the performance of several anthropometric measures in predicting clusters of cardiovascular risk factors in an adult population in Galicia (NW Spain), with adjustment for age and gender.

KEY WORDS: ROC curve; generalized additive models; bootstrap; cardiovascular risk factors; anthropometric measures.

1 INTRODUCTION

The discriminatory capacity of a continuous marker or diagnostic test Y, is usually measured by means of the receiver operating characteristic (ROC) curve [1, 2]. Under the conventional assumption that high marker values are indicative of disease, classification on the basis of Y of an individual as healthy (D) or diseased ($\bar{D}$) can be made by the choice of a cut-off value c, such that, if $Y \geq c$, the individual is classified as diseased, and if $Y < c$, the individual is classified as healthy. Hence, each cut-off value c chosen will give rise to a true positive fraction, $TPF(c) = P[Y \geq c|D]$, and a false positive fraction, $FPF(c) = P\left[Y \geq c|\bar{D}\right]$. In such a situation, the ROC curve is defined as the set of all TPF-FPF pairs that can be obtained on the cut-off value c varying, $\{(TPF(c), FPF(c)), c \in (-\infty, \infty)\}$, or, equivalently, as the function of the form $ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right)$ for $t \in (0,1)$, where S_D and $S_{\bar{D}}$ denotes the survival functions of Y in diseased and healthy subjects, respectively.

In many practical situations, however, a marker's discriminatory capacity may be affected by a set of continuous and/or categorical covariates $\mathbf{X}$. This is the case of our endocrine study, in which the performance of each anthropometric measure in detecting clusters of cardiovascular risk factors changes with individuals' age and gender (see Section 4 for details). The ROC curve may thus be of little value if important covariates are omitted. Moreover, in such situations, interest must be focused on assessing marker Y's discriminatory capacity by reference to the values assumed by $\mathbf{X}$. Hence, if the conditional survival functions of Y_D and $Y_{\bar{D}}$, given $\mathbf{X}$, are denoted $S_{D\mathbf{X}}$ and $S_{\bar{D}\mathbf{X}}$ respectively, the conditional or covariate-specific ROC curve is defined as

$$ROC_{\mathbf{X}}(t) = S_{D\mathbf{X}}\left(S_{\bar{D}\mathbf{X}}^{-1}(t)\right), t \in (0,1). \quad (1)$$

To study the effect of covariates on the accuracy of a diagnostic test, various ROC regression methodologies have been proposed in the statistical literature, e.g., (1) 'induced' methodology [3]–

[6]; and (2) 'direct' methodology [7]–[10]. Induced ROC methodology is based on specifying a model for the diagnostic test result as a function of covariates, in both healthy and diseased populations, so that the induced covariate specific ROC can then be calculated on the basis of the two models. In contrast to this, in direct methodology the effect of the covariates is directly evaluated on the ROC curve. Using this methodology, the general form of the covariate-specific ROC curve is given by the following generalized linear model (GLM)

$$ROC_{\mathbf{X}}(t) = g\left(\mathbf{X}'\boldsymbol{\beta} + h(t)\right), t \in (0,1), \tag{2}$$

where $\mathbf{X}$ is a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters, h is an unknown monotone increasing function of the FPFs, and g is a known link function, describing the functional relationship between the ROC curve and the covariates. Models such as (2) define the so-called class of ROC-GLMs [11]. As regard the function h, while it is assumed to have a parametric form by Alonzo and Pepe [8], in the case of Cai and Pepe [9] and Cai [10] it remains completely unspecified.

Along with the advantage of directly evaluating the effect of the covariate on the ROC curve, direct methodology has some other appealing features, including the following: the ROC's property of being invariant to monotonic transformation of the test result is preserved; and, any possible interaction between covariates and FPFs is easy to incorporate into the regression model.

In some circumstances, however, the ROC-GLM regression model (2) can be very restrictive, since it assumes linearity in the effect of the continuous covariates. Misleading conclusions may be drawn if this effect is incorrectly specified. This restriction can be avoided by, say, using an extension of (2) as an generalized additive model (GAM) [12]. In this setting, the researcher, rather than assuming a parametric form for the effects of the continuous covariates, solely assumes that these effects could be represented by arbitrary smooth functions. The ROC-GAM can thus be expressed as:

$$ROC_{\mathbf{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h(t)\right), t \in (0,1), \tag{3}$$

where f_j and h are assumed to be smooth and unknown functions.

To date, some attempts have been made to include covariates in ROC analysis in the nonparametric field. Recently, López-de-Ullibarri et al. [13] studied the effect of covariates on the ROC curve, using a nonparametric approach based on local linear estimation of conditional survival functions in healthy and diseased subjects. In the induced ROC regression context, González-Manteiga et al. [14] and Rodríguez-Álvarez et al. [15] have proposed new nonparametric estimators of the covariate-specific ROC curve, based on kernel-type regression estimators. Although all these approaches have proved useful, they suffer from the limitation of only being able to address a single continuous covariate.

To the best of our knowledge, there have been no approaches proposed in the literature, based on direct methodology, in which the effect of continuous covariates on the ROC curve are modeled nonparametrically. The main goal of this paper is to present a new flexible estimator of the covariate-specific ROC curve based on the ROC-GAM regression model given in (3). To this end, use is made of the local scoring algorithm with an inner backfitting loop [12, 16], based on local linear kernel smoothers [17, 18]. Among the advantages of using such smoothers is the possible use of binning type acceleration techniques [19] to reduce computational time and so ensure that the problem can be adequately addressed in practical situations.

It should be noted that the class of ROC-GAM regression models given in (3) includes, by way of specific examples, other models previously addressed in the literature. For instance, if $g = \Phi$ and $h(t) = \alpha_0 + \alpha_1 \Phi^{-1}(t)$, the ROC curve follows the classic binormal model. If the effects of the covariates X_j is linear, however, then the corresponding partial functions can be expressed parametrically $f_j(X_j) = \beta_j X_j$, yielding the ROC-GLM regression model (2).

The remainder of the paper is structured as follows: Section 2 introduces the proposed algorithm for the estimation of the ROC-GAM regression model (3). This algorithm is an extension to the nonparametric framework of the algorithm proposed by Alonzo and Pepe [8]. In Section 3, the performance of the estimation procedure is evaluated by means of simulations. In Section 4, we illustrate our method using data from the endocrine field, and conclude with a discussion in Section 5. Some technical details have been added by way of an appendix.

2 ROC-GAM ESTIMATION PROCEDURE

This section presents an algorithm for the estimation of the ROC-GAM regression model (3). It should be noted that, in order to guarantee the identification of model (3), we introduce a constant α into the model and require a zero mean $E(f_j) = 0$ for the partial functions (see [12]). For study purposes, deem $\mathbf{X}$ to be a p-dimensional covariate vector and let $\left\{\left(y_i^{\bar{D}}, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(y_j^{D}, \mathbf{x}_j^{D}\right)\right\}_{j=1}^{n_D}$ be two independent random samples drawn from the healthy and diseased populations, respectively.

Unlike standard regression analysis, in ROC-GAM regression model (3) the dependent variable is not directly observable. This makes it necessary for another interpretation to be given to the ROC curve (see [7], [20]). The key idea for fitting the model (3) is based on the placement values [21] of Y_D defined as $PV_D \equiv S_{\bar{D}\mathbf{X}}(Y_D)$. Given that

$$E\left[I\left[PV_D \leq t\right] | \mathbf{X}\right] = ROC_{\mathbf{X}}(t),$$

the covariate-specific ROC curve can be viewed as the conditional expectation of the binary variable $B_{Dt} = I\left[PV_D \leq t\right]$. The ROC-GAM regression model (3) can therefore be viewed as a regression model for B_{Dt}. This suggests [8] that estimation of the ROC-GAM regression model (3) can be based on the following algorithm:

1. choose a set $T = \{t_l, l = 1, \ldots, n_T\}$ of FPFs;
2. estimate $S_{\bar{D}\mathbf{X}}$ on the basis of $\left\{\left(y_i^{\bar{D}}, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$;
3. for each disease observation, calculate the estimated placement value $PV_j = \hat{S}_{\bar{D}\mathbf{x}_j^D}\left(y_j^D\right), j = 1, \ldots, n_D$;
4. for all $t_l \in T$ and each disease observation, calculate the binary placement value indicator $\hat{B}_{jt_l} = I\left[PV_j \leq t_l\right], l = 1, \ldots, n_T, j = 1, \ldots, n_D$; and
5. fit the following ROC-GAM binary regression model

$$ROC_{\mathbf{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h(t)\right), \tag{4}$$

to the data $\left\{\left(\hat{B}_{jt_l}, \left\{\mathbf{x}_j^D, t_l\right\}\right), l = 1, \ldots, n_T, j = 1, \ldots, n_D\right\}$, and obtain the estimates $\widehat{ROC}_{\mathbf{X}}(t)$.

To estimate the binary GAM (4), the local scoring algorithm with backfitting was used [16, 22]. Briefly, the local scoring algorithm is analogous to the use of iterative reweighted least squares [23] for solving nonlinear regression equations. The backfitting algorithm cycles through each of the covariates $X_j, (j = 1, \ldots, p)$ and the estimates $\hat{f}_j$ are obtained by applying local linear kernel smoothers [17, 18] to the corresponding partial residuals. We have chosen to use of local linear kernel smoothers, since these estimators enable us to implement binning [19] directly into the estimation algorithm and computing time is thus drastically reduced (see Appendix for a detailed description).

It should be noted that the observations used to fit the binary GAM (4) are no longer independent, since each disease observation is 'compared' with all $t_l \in T$. It is well known that in the presence of correlated errors, standard bandwidths selectors fail to work and can result in an over (or under) fit (see e.g. [24]). In this study, the cross-validation criterion was used for the automatic choice of bandwidths. As pointed out, this choice of bandwidths may be far from optimal. Nevertheless, the estimation procedure seems to perform reasonably well in the simulation studies presented in Section 3 below. Finally, to obtain an increasing monotone estimation curve of h, we used the monotone smoothing technique proposed by Friedman and Tibshirani [25].

To implement the estimation procedure presented at the beginning of this section, the conditional survival function in healthy subjects must be estimated, $S_{\bar{D}\mathbf{X}}$ (see Step 2). In this paper, we propose to model the effect of covariates on $Y_{\bar{D}}$ by a nonparametric location-scale regression model, such that

$$Y_{\bar{D}} = \mu_{\bar{D}}(\mathbf{X}) + \sigma_{\bar{D}}(\mathbf{X})\varepsilon_{\bar{D}}, \tag{5}$$

where $\mu_{\bar{D}}$ and $\sigma^2_{\bar{D}}$ are the regression and the variance functions, respectively, and error $\varepsilon_{\bar{D}}$ is assumed to be independent of the covariates $\mathbf{X}$, with zero mean, unit variance and survival function S. With this configuration, it can be shown that

$$S_{\bar{D}\mathbf{X}}(y) = S\left(\frac{y - \mu_{\bar{D}}(\mathbf{X})}{\sigma_{\bar{D}}(\mathbf{X})}\right).$$

Our proposal is based on using local linear kernel smoothers to estimate the regression and variance functions of (5), and then estimating the survival function S on the basis of the empirical survival distribution of the standardized residuals. More specifically, an additive covariate model is assumed, and $\mu_{\bar{D}}$ is then estimated based on the sample $\left\{\left(y_i^{\bar{D}}, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$ with local linear kernel smoothers and the backfitting algorithm being used for the unweighted case (see Appendix for details). Next, to estimate $\sigma^2_{\bar{D}}$, an additive smooth model can be fitted to the sample $\left\{\left(\left(y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)\right)^2, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$. However, to guarantee $\hat{\sigma}^2_{\bar{D}} > 0$, the transformed variance, $\log\left(\sigma^2_{\bar{D}}\right)$, is first estimated by fitting an additive covariate model (using local linear kernel smoothers, as before) to the transformed data $\left\{\left(\log\left(y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)\right)^2, \mathbf{x}_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$. The final estimate of $\sigma^2_{\bar{D}}$ can be obtained as $\exp\left(\widehat{\log\left(\sigma^2_{\bar{D}}\right)}\right)$. Nevertheless, this naïve approach may or may not work well (see [12], p. 194). Accordingly, an improved approach is used here, which consists of obtaining the estimates $\hat{\sigma}^2_{\bar{D}}$ by smoothing the data $\left\{\left(y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)\right)^2\right\}_{i=1}^{n_{\bar{D}}}$ against $\left\{\log\left(\widehat{\sigma^2_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)}\right)\right\}_{i=1}^{n_{\bar{D}}}$, with the Nadaraya-Watson estimator being used for the purpose estimator [26, 27]. Finally, the survival function S is estimated on the basis of the empirical distribution of the standardized residuals

$$\hat{S}(z) = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} I\left[\frac{y_i^{\bar{D}} - \hat{\mu}_{\bar{D}}(\mathbf{x}_i^{\bar{D}})}{\hat{\sigma}_{\bar{D}}(\mathbf{x}_i^{\bar{D}})} \geq z\right].$$

3 SIMULATION STUDY

This section reports the results of a simulation study conducted to study the practical behavior of the estimation procedure described in Section 2 above.

In step 5 of our estimation procedure, the assumption that the observations are independent no longer holds, and so this study first examined the behavior of the cross-validation criterion used to choose the optimal smoothing parameters. To this end, the performance of the proposed method was compared against that of the induced nonparametric regression approach [15], in which this particular issue is of no concern. In addition, the efficiency and robustness of our model was studied, by comparing it to the direct parametric approach proposed by Alonzo and Pepe [8] and the semiparametric model suggested by Cai [10].

Data were simulated from two scenarios, namely,

- Scenario I

$$Y_D = 2 + 4X_1 + \varepsilon_D,$$
$$Y_{\bar{D}} = 1.5 + 3X_1 + 0.5\varepsilon_{\bar{D}},$$

with this configuration, the corresponding covariate-specific $ROC_X(t)$ is

$$ROC_X(t) = \Phi\left(0.5 + X_1 + 0.5\Phi^{-1}(t)\right).$$

- Scenario II

$$Y_D = 1 + \sin(\pi(X_1 + 2)) + (.25 + .25(X_1 + 1))\,\varepsilon_D,$$
$$Y_{\bar{D}} = 0.5\exp(X_1) + (.25 + .25(X_1 + 1))\,\varepsilon_{\bar{D}},$$

with this configuration, the corresponding covariate-specific $ROC_X(t)$ is

$$ROC_X(t) = \Phi\left(\frac{1 + \sin(\pi(X_1 + 2)) - 0.5\exp(X_1)}{(.25 + .25(X_1 + 1))} + \Phi^{-1}(t)\right).$$

In both scenarios, X_1 is uniformly distributed on $[-1, 1]$, $\varepsilon_{\bar{D}}$ and ε_D have the standard normal distribution, and Φ denotes the CDF of a standard normal variable.

The discrepancy between the estimator of the covariate-specific ROC curve and the true ROC curve was measured in terms of the empirical version of the global mean squared error (MSE):

$$MSE = \frac{1}{n_{X_1}} \sum_{l=1}^{n_{X_1}} \frac{1}{n_\tau} \sum_{r=1}^{n_\tau} \left(\widehat{ROC}_{X_1=x_l}(t_r) - ROC_{X_1=x_l}(t_r)\right)^2,$$

with $x_l = -1.0 + 2\frac{l-1}{n_{X_1}-1}$, $l = 1, \ldots, n_{X_1}$, $t_r = \frac{r-1}{n_\tau - 1}$, $r = 1, \ldots, n_\tau$ and n_{X_1}=n_τ=50.

It should be noted that to fit the ROC-GAM model (3), the set T of FPFs must be chosen

(step 1 of the estimation algorithm). In the results shown below, $n_T = 50$ and equally-spaced values were considered. Other simulations were performed for different cardinalities of set T, with very similar results (not shown). Table 1 lists the averages and standard deviations of the MSEs obtained in 1000 data sets simulated from Scenario I and II. In all cases, the same sample size was considered for healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 50, 100, 200, 500$. As expected, the MSE decreases as the sample sizes increase. As can be seen, the proposed ROC-GAM regression model performs similarly to the induced nonparametric (Induced NP) approach, which suggests that the behavior of the cross-validation criterion used to choose the optimal smoothing parameters is satisfactory. As would be expected under Scenario I, the models that displayed the best performance were the parametric and semi-parametric approaches, though our method also performed satisfactorily (see Figure 1(a)). For Scenario II, however, the effect of the covariate on the ROC curve was far from linear, and the estimates obtained by the parametric and semi-parametric models were thus not suitable, as can be seen from Figure 1(b). The good performance of the proposed model is clearly observable in Figure 2, where the true ROC curve (solid line) and the corresponding average of the estimated ROCs (dashed line) for Scenario II are shown. The estimated standard deviation is shown in Figure 3. Figure 4 depicts the average of the AUCs estimated by the proposed model, along with the 2.5 and 97.5 simulation quantiles, for the different sample sizes. As can be seen, the proposed estimator behaves well, with variance decreasing as sample size increases.

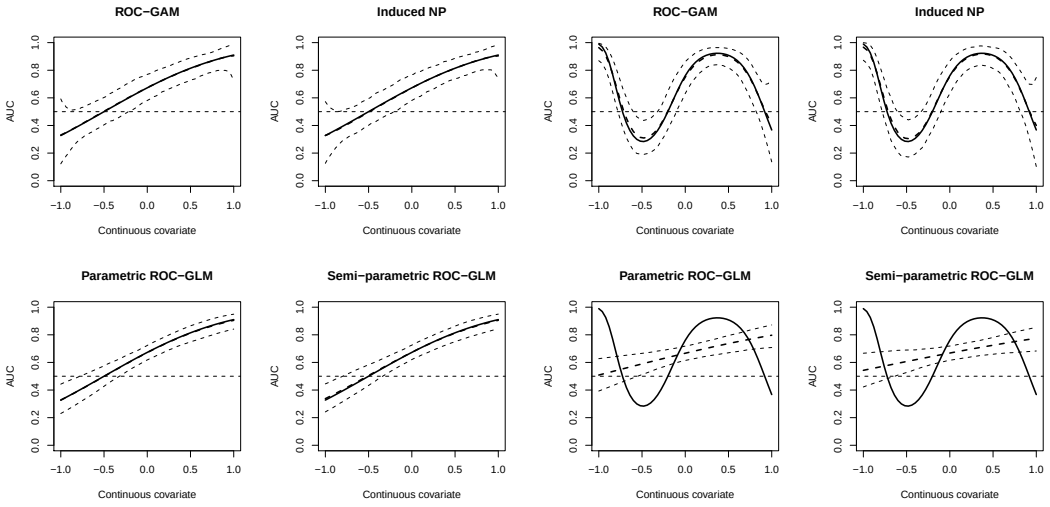
Lastly, the behavior of the proposed model was examined in a scenario with more than one continuous covariate. We have simulated data from the following scenario:

- Scenario III

$$Y_D = .5 \sin(\pi(X_1 + 1)) + .5 \exp(X_1) - X_2^2 + .5\varepsilon_D,$$
$$Y_{\bar{D}} = .5 \exp(X_1) - 2X_2^2 + .5\varepsilon_{\bar{D}},$$

Table 1: Average and standard deviation of estimated mean squared error (MSE) (x 1000) based on the 1000 simulated data sets, yielded by the proposed ROC-GAM, the induced nonparametric approach (Induced NP), Alonzo-Pepe's direct parametric approach (Parametric ROC-GLM) and Cai's direct semi-parametric approach (Semi-parametric ROC-GLM)

		Sample size			
Scenario	Model	50	100	200	500
I	ROC-GAM	17.454 (11.224)	9.393 (5.105)	5.777 (2.306)	3.508 (0.921)
	Induced NP	17.647 (10.054)	9.311 (4.700)	5.163 (2.177)	2.579 (0.894)
	Parametric ROC-GLM	6.125 (5.828)	2.979 (2.945)	1.511 (1.337)	0.595 (0.559)
	Semi-parametric ROC-GLM	9.586 (6.079)	5.596 (3.250)	3.358 (1.486)	1.864 (0.727)
II	ROC-GAM	32.532 (15.031)	15.887 (6.522)	7.706 (3.160)	3.293 (1.212)
	Induced NP	38.923 (17.904)	18.609 (7.832)	8.811 (3.740)	3.454(1.385)
	Parametric ROC-GLM	64.705 (7.108)	64.089 (4.876)	62.164 (2.759)	61.332(1.628)
	Semi-parametric ROC-GLM	66.414 (7.000)	63.631 (3.971)	61.030 (1.873)	59.618(0.832)



(a) ScenarioI

(b) ScenarioII

Figure 1: True AUC (solid line) versus the average of simulated AUCs (dashed line), along with the 2.5 and 97.5 simulation quantiles, for $n = 200$, based on 1000 estimates. From left to right: Proposed ROC-GAM, induced nonparametric (Induced NP) approach, Alonzo-Pepe's direct parametric approach (Parametric ROC-GLM), Cai's direct semi-parametric approach (Semi-parametric ROC-GLM).

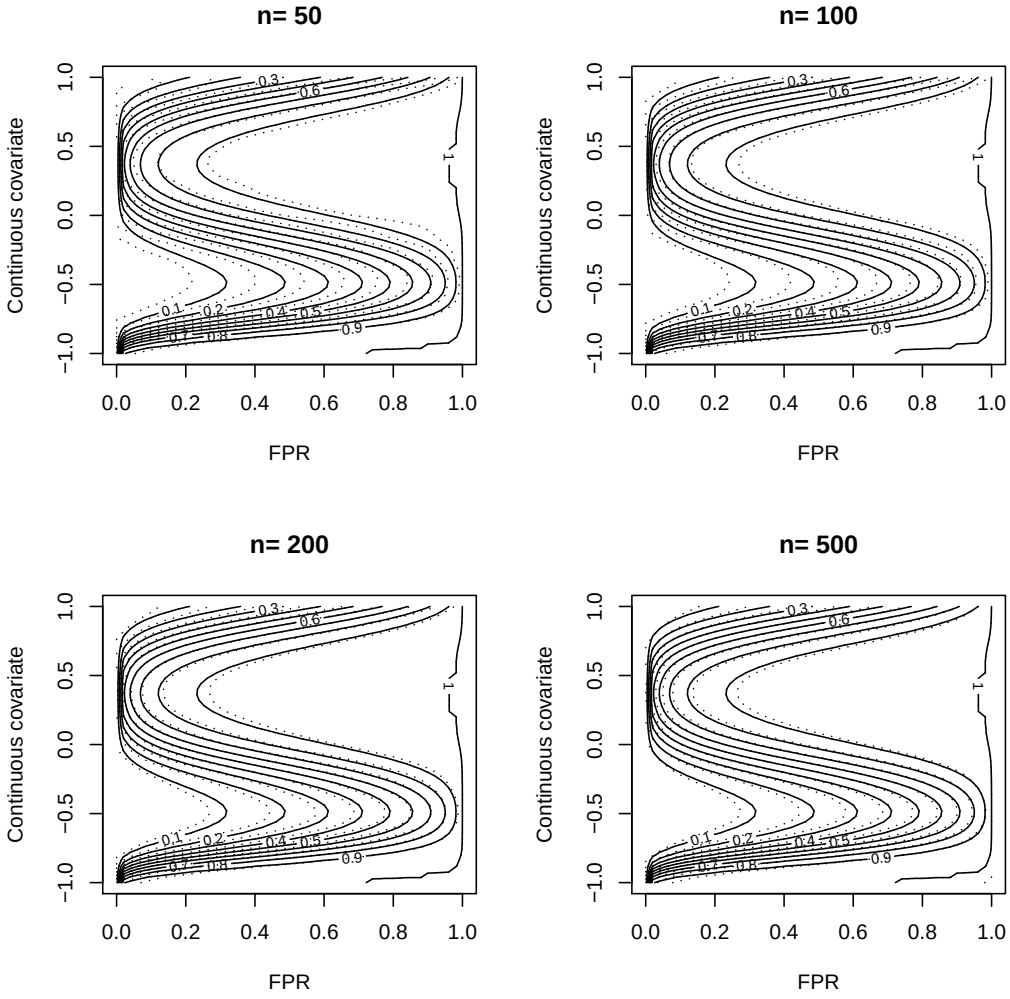


Figure 2: Contour plot for the true ROC (solid line) versus the average of simulated ROCs (dashed line) yielded by the ROC-GAM, for Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

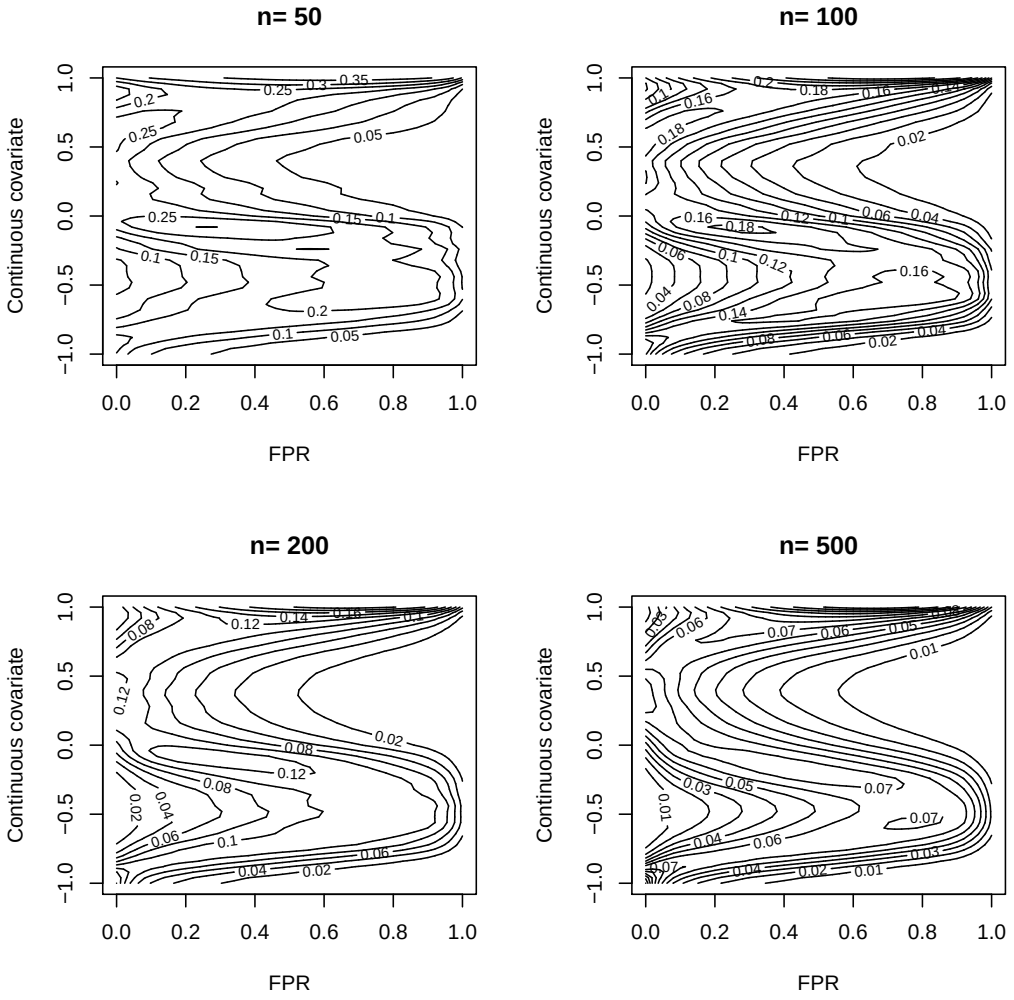


Figure 3: Contour plot for standard deviation of simulated ROCs yielded by the ROC-GAM, for Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

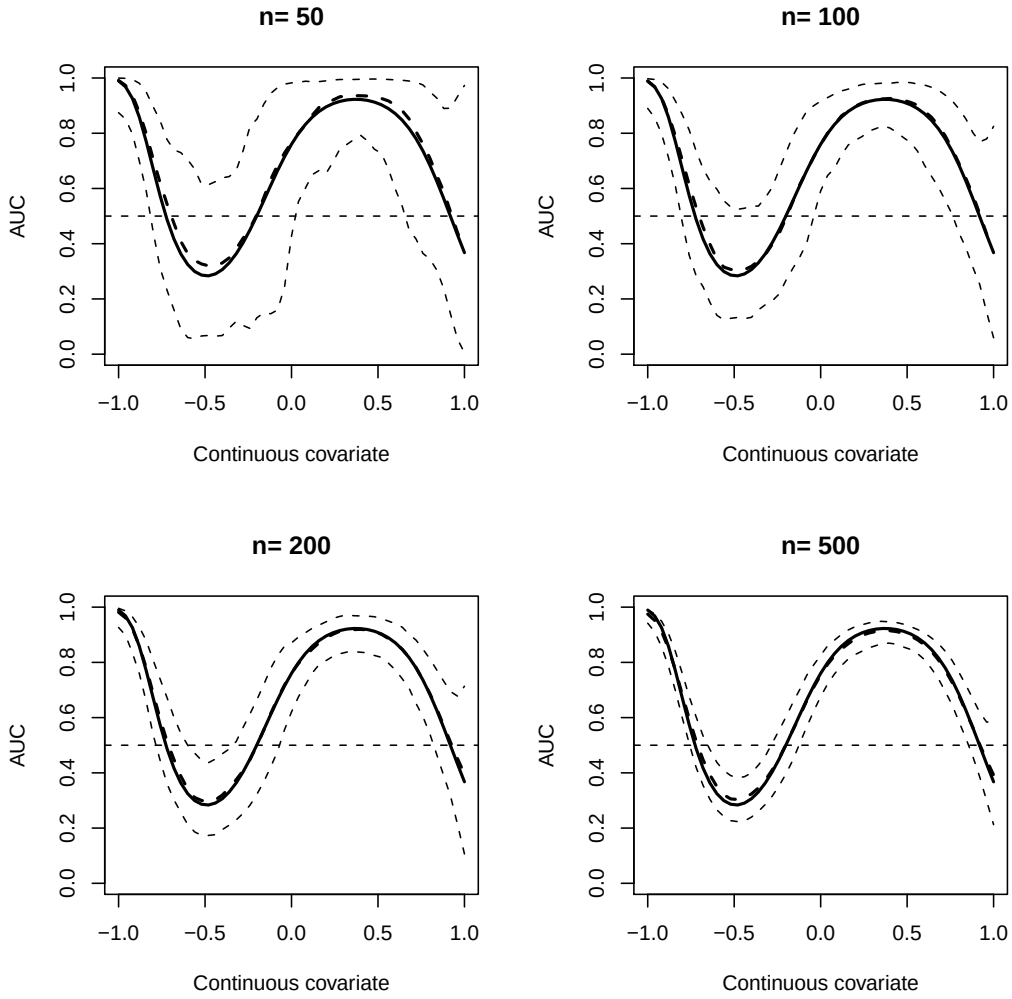


Figure 4: True AUC (solid line) versus the average of simulated AUCs (dashed line) yielded by the ROC-GAM, along with 2.5 and 97.5 simulation quantiles for the ROC-GAM, in Scenario II and for different samples sizes ($n = n_D = n_{\bar{D}}$), based on 1000 estimates.

with this configuration, the corresponding covariate-specific $ROC_X(t)$ is

$$ROC_X(t) = \Phi\left(\sin\left(\pi(X_1+1)\right) + 2X_2^2 + \Phi^{-1}(t)\right),$$

where X_1 and X_2 are uniformly distributed on $[-1,1]$, $\varepsilon_{\bar{D}}$ and ε_D have the standard normal distribution, and Φ denotes the CDF of a standard normal variable.

For this particular scenario, we investigated the performance of the ROC-GAM when estimating the two partial functions $f_1(X_1) = \sin(\pi(X_1+1))$ and $f_2(X_2) = 2X_2^2$, and the baseline function $h(t) = \Phi^{-1}(t)$. Averages of the results are graphically depicted in Figure 5. The same sample size was considered for both healthy and diseased subjects, with $n = n_D = n_{\bar{D}} = 400$. The good performance of the resulting estimates is evident, with the functional forms of the corresponding true curves being recovered very successfully.

4 APPLICATION TO REAL DATA

Coronary heart and cerebrovascular disease are one of the leading causes of adult mortality worldwide. There are several cardiovascular disease (CVD) risk factors for developing these vascular diseases, including diabetes mellitus, dyslipidemia, arterial hypertension and obesity. These CVD risks have a metabolic basis and the common characteristic of tending to occur in clusters (e.g., metabolic syndrome), in that when one risk factor is present in any subject, another is also likely to appear in the same individual. Although the metabolic defect ultimately responsible for such clustering is not known, obesity is nonetheless well documented as being a common denominator [29].

Yet, serious concern exists as to which anthropometric measure related to excess body fat best predicts cardiovascular risks. Although there is some evidence to indicate that indices of abdominal fat accumulation, such as waist circumference (WC) or waist-to-hip ratio (WHR), may be better predictors than body mass index (BMI) [30, 31], the question of which obesity index is best for predicting CVD risks is still a matter of controversy in biomedical research [32, 33].

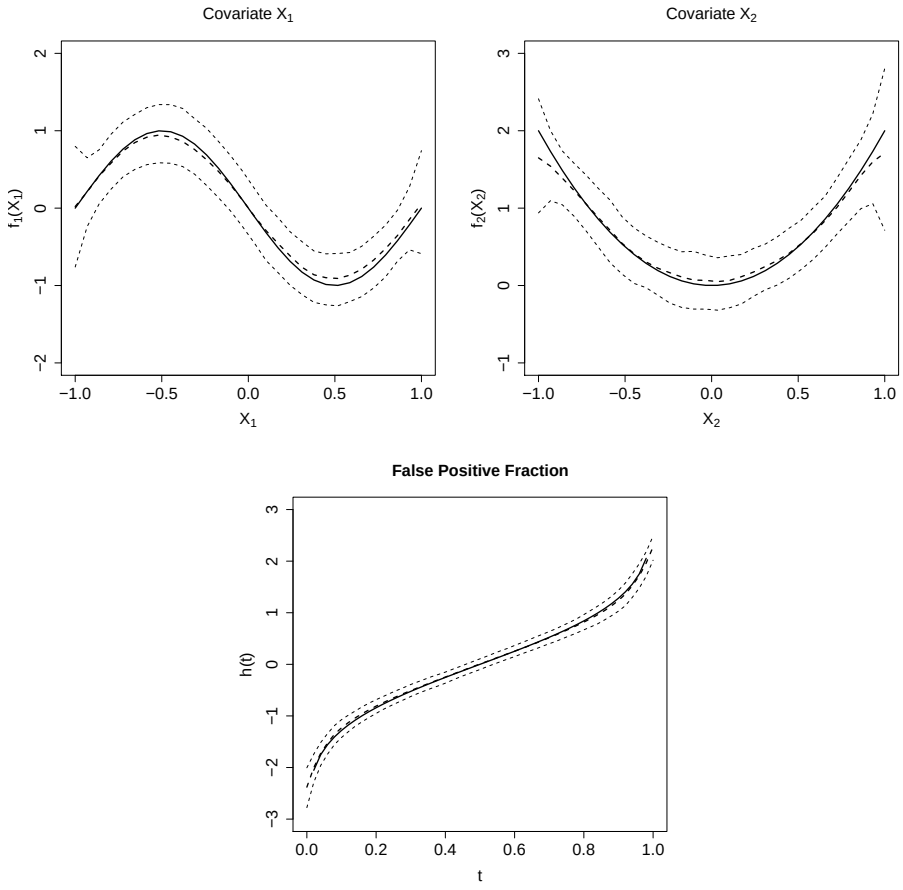


Figure 5: Simulation results based on 1000 replicated samples obtained under Scenario III. From left to right: true curve f_1 (solid line) and average estimate $\hat{f}_1$ (dashed line); true curve f_2 (solid line) and average estimate $\hat{f}_2$(dashed line); and true curve h (solid line) and average estimate $\hat{h}$ (dashed line). In all cases, the 2.5 and 97.5 simulation quantiles have also been plotted.

In view of the existing gaps in knowledge, we applied the proposed ROC-GAM methodology to an endocrine study [34, 35], with the aim of assessing the performance of BMI, WC and WHR for predicting clusters of cardiovascular risk factors in an age- and gender-adjusted adult population in Galicia (NW Spain). Since it is well established that anthropometric measures perform differently according to gender, the age-by-gender interaction was included in the ROC-GAM regression models.

4.1 Data source

This study consisted of field work covering a random sample, representative of the Galician adult population (2850 subjects, age range 18-85 years). Direct anthropometric measurements were taken, including weight (in kilograms), height (in meters), WC (in centimeters) and hip circumference (in centimeters), and the WHR was subsequently calculated. BMI was computed as weight divided by height squared. A diseased subject was defined as any person presenting with two or more CVD risk factors (raised triglycerides, reduced HDL-cholesterol, raised blood pressure and raised fasting plasma glucose) as per the International Diabetes Federation criteria [36]. Of the total of 2850 subjects, 46.2% were men (899 healthy and 418 diseased) and 53.8% women (1250 healthy and 273 diseased). A detailed description of this dataset can be found in [35].

4.2 Statistical analysis

For each anthropometric measure, separate covariate-specific ROC curves were calculated, assuming the following multivariate interaction model:

$$ROC_{(Age,Gender)} = \Phi\left(\beta_0 + f_{Men}(Age)\mathbf{1}_{\{Gender=Men\}} + f_{Women}(Age)\mathbf{1}_{\{Gender=Women\}} + h(t)\right),$$

where f_{Men} and f_{Women} are smooth functions of Age in men and women, respectively, $h(t)$ is a smooth function of the false positive fraction, and $\mathbf{1}_A$ denotes the indicator function of event A.

In addition to the estimated covariate-specific ROC curves, another summary measure of accuracy, the AUC, was obtained. Bootstrap-based methods were used for constructing confidence inter-

vals for this measure [28]. For our purposes, the global sample will be denoted as $\{(\mathbf{x}_k, y_k, d_k)\}_{k=1}^{n_{\bar{D}}+n_D}$, where d is a binary indicator, taking the value 1 for disased and 0 for healthy individuals.

Given a point $\mathbf{x}$ in the range of $\mathbf{X}$ the steps for the construction of the confidence interval were as follows:

1. for $b = 1, \ldots, 500$, draw a random sample of size $n_{\bar{D}} + n_D$, with replacement from the global population; and,
2. from $\left\{\left(\mathbf{x}_k^b, y_k^b, d_k^b\right)\right\}_{k=1}^{n_{\bar{D}}+n_D}$ obtain $\widehat{ROC}^b_{\mathbf{X}=\mathbf{x}}(t)$ and compute

$$\widehat{AUC}^b_{\mathbf{X}=\mathbf{x}} = \int_0^1 \widehat{ROC}^b_{\mathbf{X}=\mathbf{x}}(t)dt.$$

Once the above process has been completed, the 100 per cent x $(1-\alpha)$ limits for the confidence interval for the true $AUC_{X=x}$ are given by

$$\left(\widehat{AUC}^{\alpha/2}_{\mathbf{X}=\mathbf{x}}, \widehat{AUC}^{1-\alpha/2}_{\mathbf{X}=\mathbf{x}}\right)$$

where $\widehat{AUC}^p_{\mathbf{X}=\mathbf{x}}$ represents the p-percentile of the estimated $\widehat{AUC}^b_{\mathbf{X}=\mathbf{x}}$ $(b = 1, \ldots, B)$.

4.3 Results

The covariate-specific ROC curves are plotted in Figure 6. Figure 7 depicts the covariate-specific AUCs together with the corresponding 95% bootstrap-confidence intervals. As can be seen from these figures, the discriminatory capacity of the measures studied was affected by sex and age alike. In general, age displayed a more marked effect among women than among men, particularly in the case of BMI and WC. These results suggest the need for the incorporation of the age-gender interaction in the ROC-GAM regression models fitted.

In the case of men, the three measures will be seen to behave similarly, with accuracy tending to decline progressively and eventually starting to lose significance around age 55 years in the case of WC, and age 65 years in the case of BMI and WHR.

With respect to women, the AUCs indicate very good discriminatory capacity for the youngest women, especially in the case of BMI and the WC, with values greater than 0.8. This better behavior of BMI and WC vis-à-vis WHR remains constant until the age of 45 years. From here onwards, the three measures reach a plateau until subjects are in their sixties. Thereafter, WHR becomes the best predictor. From the age of sixty, both BMI and the WC decline progressively and start losing significance from age $65 - 70$ years onwards. WHR, on the other hand, retains its discriminatory capacity beyond the age of 70 years, and only then does its discriminatory capacity disappear.

It is interesting to observe the similar behavior displayed by BMI and WC as compared to WHR, particularly in the case of women. As will be observed, WHR's discriminatory capacity displays a more constant behavior than do the other two measures over subjects' lifetime, with no major differences between men and women in this instance.

5 DISCUSSION

This paper presents a new nonparametric estimator for covariate-specific ROC curve based on direct modeling. The proposed estimator enables a set of continuous and/or categorical covariates, and any possible interactions, to be incorporated into the regression model for the ROC curve.

The proposed estimation procedure is based on a combination of local scoring and backfitting algorithms, and uses local linear kernel smoothers. Use is made of the cross-validation criterion to choose optimal smoothing parameters, and binning techniques to speed up computation time. The process of estimation of the ROC-GAM regression model requires the conditional survival function in the healthy population to be estimated. This paper proposes a location-scale regression model for the test result in the healthy population. Local linear estimators are used to estimate regression and variance functions, and the survival function is empirically estimated on the basis of standardized residuals.

The steps of the proposed estimation algorithm imply the non-independence of the observation upon which the estimation of the ROC-GAM regression model is based. Despite this, the results of

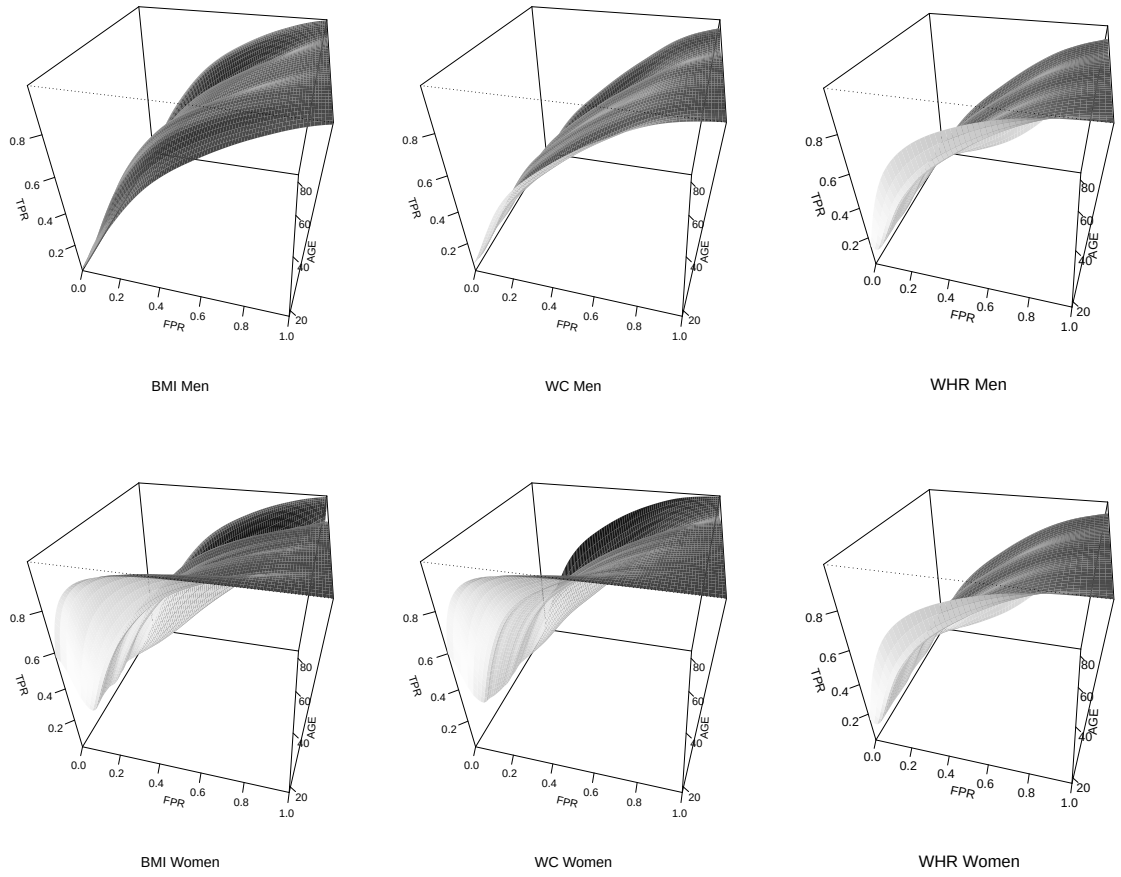


Figure 6: ROC surfaces, age and gender, for different anthropometric biomarkers. Top row (for Men): BMI, WC and WHR. Bottom row (for Women): BMI, WC and WHR.

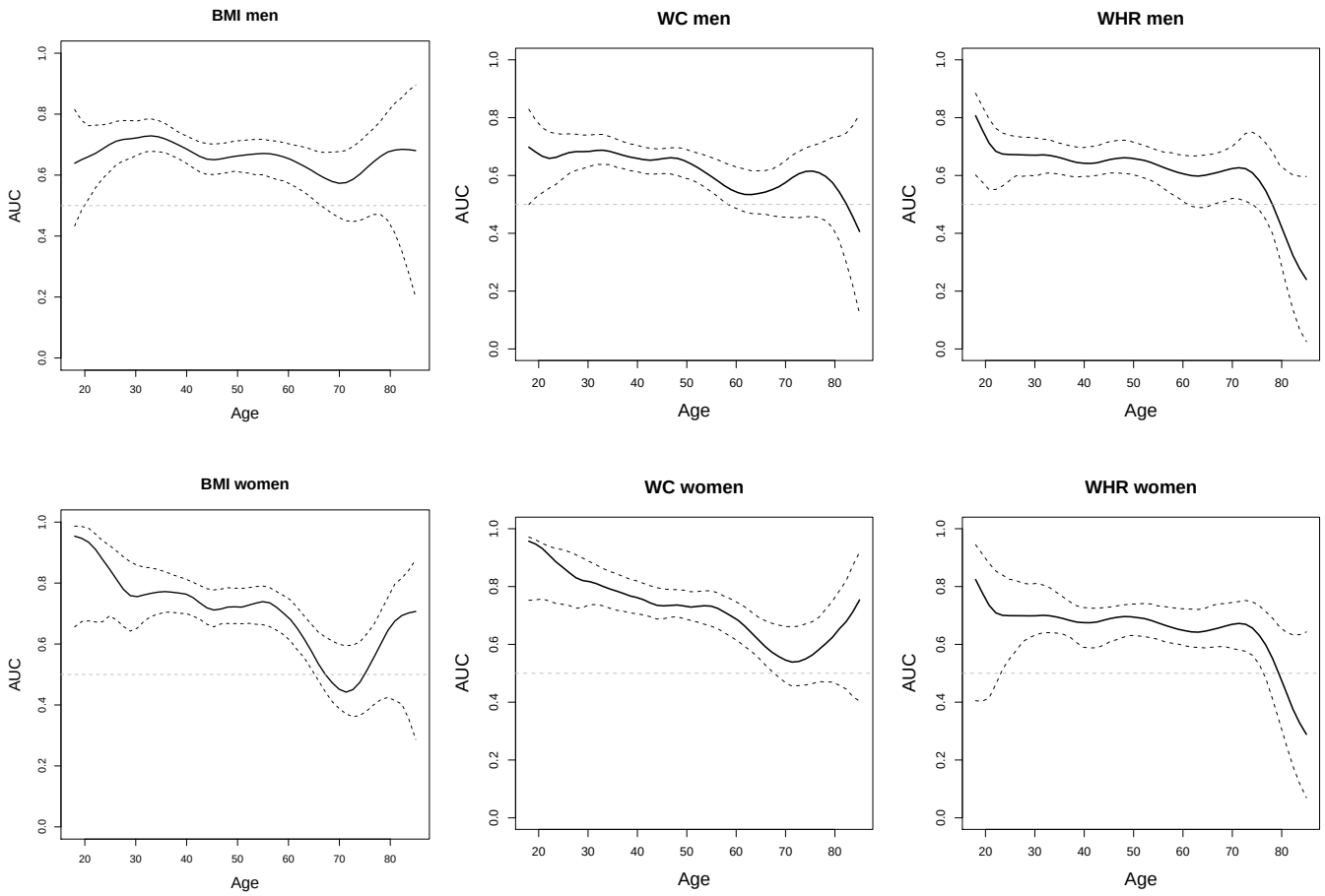


Figure 7: AUCs adjusted by age and gender with 95% bootstrap confidence bands for BMI, WC and WHR.

the simulation study that compares the new and induced nonparametric approaches would appear to indicate the good performance of our model, and, specifically, that of the cross-validation criterion used for automatic selection of the smoothing parameters. We also compared the performance of the ROC-GAM against previous direct approaches, as well as its behavior in the presence of more than one continuous covariate. The results show that our estimator also performs well in all these situations.

The methodology proposed in this paper was used to analyze an endocrine dataset, in order to evaluate the effect of age and gender on the discriminatory capacity of different anthropometric measures when it came to detecting the presence of two or more cardiovascular risk factors. The results show the different behavior of the measures considered according to individuals' sex and age, with the latter aspect proving particularly evident in the case of women. Even so, our study does not suggest that any one measure studied has a discriminatory capacity that is clearly better than that of the others over the span of a subject's lifetime. It should be noted that, the results obtained in this study are representative of Galicia alone and so cannot be extrapolated to other populations.

Although our methodology was motivated by endocrine data, it could nonetheless be applied in any other biomedical field where the accuracy of a given diagnostic test might depend on a set of continuous and factor-type covariates.

This paper has focused on the presentation of the process of estimation of the ROC-GAM model, yet considerably more work is still needed. For instance, an interesting field for future research would be the development of test statistics for the effects of the covariates on the ROC curve, the linearity of the effects of the continuous covariates, or the existence of interactions among covariates, aspects on which there is still little in the literature. Special attention should be paid to Cai and Zheng's paper [37], in which different procedures to check the goodness-of-fit for some parametric and semiparametric ROC regression approaches are proposed.

In our endocrine study, different ROC-GAM regression models were separately fitted for each anthropometric measure. It would, however, be desirable for all the markers to be incorporated into a single regression model. Having more than one observation per subject would mean that, in such

a situation, there would be a new source of non-artificial correlation in the data. Consequently, an interesting issue for future research would be to incorporate a structure of correlation among observations pertaining to the same individual into the process of estimation of the ROC-GAM regression model.

It should be pointed out that, in practical situations, the new methodology may represent a flexible exploratory tool for identifying non-linear covariate effects on the ROC curve. ROC-GAMs can also be used in a diagnostic mode, as an aid for choosing parametric transformation for the covariates (where necessary), in which case a ROC-GLM could thus be used.

Software implementing the ROC-GAM estimation procedure proposed in this paper can be obtained by contacting the first-mentioned author at mariajose.rodriguez.alvarez@usc.es.

Acknowledgements

The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and PGIDIT07PXIB300191PR, and would also like to thank the Galician Endocrinology & Nutrition Foundation (*Fundacion de Endocrinoloxia e Nutricion Galega - FENGA*) for having supplied the database used in this study.

Appendix: Computational aspects

The local scoring algorithm

Given a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ of $(Y, \mathbf{X})$, the steps of the local scoring algorithm for estimating a GAM for binary response

$$\mu = E[Y|\mathbf{X}] = g\left(\alpha + f_1(X_1) + \ldots + f_p(X_p)\right),$$

are as follows:

Initialise. Compute the initial estimates, $\hat{\alpha} = g^{-1}(\bar{y})$, $\hat{f}_j^0 = 0$ $(1 \leq j \leq p)$ and $\hat{\mu}_i^0 = \bar{y} = n^{-1}\sum_{i=1}^n y_i$ $(i = 1, \ldots, n)$.

Step 1. Form the adjusted dependent variables $\tilde{y}_i = \hat{\eta}_i^0 + \left(y_i - \hat{\mu}_i^0\right)/g'\left(\eta_i^0\right)$ and weights $\tilde{w}_i = g'\left(\eta_i^0\right)^2/\hat{\mu}_i^0\left(1 - \hat{\mu}_i^0\right)$, where $\hat{\eta}_i^0 = \hat{\alpha} + \hat{f}_1^0(x_{i1}) + \ldots + \hat{f}_p^0(x_{ip})$.

Step 2. Fit an additive model to $\tilde{y}$ using the backfitting algorithm (explained in the following subsection) with weights $\tilde{w}$ and compute the updates $\hat{f}_j(x_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Step 3. Repeat Step 1, with $\hat{\mu}_i^0$ being replaced by $\hat{\mu}_i = g(\hat{\eta}_i)$ for $i = 1, \ldots, n$, until

$$\frac{\left|D\left(\hat{\mu}^{\mathbf{0}}, \mathbf{Y}\right) - D\left(\hat{\mu}, \mathbf{Y}\right)\right|}{D\left(\hat{\mu}^{\mathbf{0}}, \mathbf{Y}\right)} \leq \varepsilon,$$

where ε is a small threshold and $D(\hat{\mu}, \mathbf{Y}) = -2\sum_{i=1}^n \left[y_i \log(\hat{\mu}_i) + (1 - y_i)\log(1 - \hat{\mu}_i)\right]$.

Fitting weighted additive models

This section describes an algorithm for fitting the additive covariate model

$$E[Y|\mathbf{X}] = \alpha + f_1(X_1) + \ldots + f_p(X_p).$$

The algorithm discussed below is a modification to the weighted case of the backfitting algorithm [22]. The backfitting algorithm cycles through the covariates X_j $(j = 1, \ldots, p)$, and estimates each f_j by applying local linear kernel smoothers to the partial residuals. These residuals are obtained by removing the estimated effects of the other covariates.

Given a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ of $(Y, \mathbf{X})$ weighted by $\{w_i\}_{i=1}^n$, the steps of the estimation algorithm are as follows:

Initialise. Compute the initial estimates $\hat{\alpha} = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$ and $\hat{f}_j^0(x_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Step 1. For $j = 1, \ldots, p$ calculate residuals by removing the estimated effects of all the other

covariates:

$$y_i^j = y_i - \hat{\alpha} - \sum_{s=1}^{j-1} \hat{f}_s\left(X_{is}\right) - \sum_{s=j+1}^{p} \hat{f}_s^0\left(X_{is}\right),$$

and compute the weighted local linear kernel estimators for $i = 1, \ldots, n$ [17]

$$\hat{f}_j\left(x_{ij}\right) = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} s_j^0\left(x_{ij}\right) & s_j^1\left(x_{ij}\right) \\ s_j^1\left(x_{ij}\right) & s_j^2\left(x_{ij}\right) \end{pmatrix}^{-1} \begin{pmatrix} u_j^0\left(x_{ij}\right) \\ u_j^1\left(x_{ij}\right) \end{pmatrix} \tag{6}$$

where $s_j^r\left(x\right) = \sum_{i=1}^n \left(w_i \cdot L_j^r\left(x, x_{ij}\right)\right)$ and $u_j^r\left(x\right) = \sum_{i=1}^n \left(w_i \cdot L_j^r\left(x, x_{ij}\right) \cdot y_i^j\right)$, with

$$L_j^r\left(x, y\right) = \left(2\pi\right)^{-1/2} \left(x - y\right)^r \exp\left(-0.5\left(h_j^{-1}\left(x - y\right)\right)^2\right),$$

and with h_j being the bandwidth associated with the estimation of $\hat{f}_j$.

Step 2. Repeat **Step** 1 with $\hat{f}_j^0$ replaced by $\hat{f}_j$, until the convergence criterion

$$\sum_{i=1}^{n} \left[\hat{f}_j\left(x_{ij}\right) - \hat{f}_j^0\left(x_{ij}\right)\right]^2 \Big/ \sum_{i=1}^{n} \hat{f}_j^0\left(x_{ij}\right)^2$$

is below some small threshold ε for all the $j = 1, \ldots, p$.

In the unweighted case, the above backfitting is reduced to that introduced by Opsomer [22]. In such a case, Opsomer and Ruppert [38] (see also [39]) established the theoretical convergence and consistency of the method. The backfitting algorithm's performance in practice was revised by Nielsen and Sperlich [40].

Bandwidth selection

It is well known that the estimates obtained for the model depend heavily on the bandwidths $(h_1 \ldots, h_p)$ used in the local linear kernel estimates of the partial functions $(f_1 \ldots, f_p)$. The bandwidths are a trade-off between the bias and variance of the resulting estimates. Various proposals for an optimal selection have been suggested for the additive models, yet the difficulty of asymptotic

theory in a backfitting context means that nowadays optimal selection is still a challenging, open problem. Cross-validation was used for the automatic choice of bandwidths.

In each of the cycles of the algorithm, the bandwidth (h_j) used to obtain the estimates $\hat{f}_j$ in equation (6) was automatically selected by minimizing the following weighted cross-validation error criterion:

$$CV_j = \sum_{i=1}^{n} w_i \left(y_i^j - \hat{f}_j^{(-i)} \left(x_{ij} \right) \right)^2$$

where $\hat{f}_j^{(-i)}$ is the estimate obtained without the i^{th} element of the sample.

Cross-validation implies a high computational cost, inasmuch as it is necessary to repeat the estimation operations several times in order to select the optimal bandwidths. To speed up this process, we used binning-type acceleration techniques [17, 18] to obtain the binning approximations of $\hat{f}_j$ in each of the iterations of the estimation algorithm.

The binning approximations were obtained from the binning sample $\left\{ x_r^{\bullet j}, y_r^{\bullet j} \right\}$ and the weights $\{ w_r^{\bullet} \}$ $(1 \leq r \leq N)$, with

$$x_1^{\bullet j} < \ldots < x_N^{\bullet j}$$

being a grid of equidistant points along the j^{th} direction. Let us consider δ the distance between consecutive grids. The binning responses $y_r^{\bullet j}$ and the binning weights $W_r^{\bullet}$ are constructed according to $w_r^{\bullet} = \sum_{i=1}^{n} w_r^{\bullet i}$ and $y_r^{\bullet j} = \sum_{i=1}^{n} w_r^{\bullet i} y_i^j$ with

$$w_r^{\bullet i} = w_i \left(1 - \left| x_{ij} - x_r^{\bullet j} \right| / \delta \right)_+$$

The binning approximation of the estimator $\hat{f}_j \left(x \right)$ is obtained by applying the approximations

$$s_j^r \left(x \right) \approx \sum_{l=1}^{N} L_j^r \left(x, x_l^{\bullet j}; h_j \right) w_l^{\bullet} \text{ and } t_j^r \left(x \right) \approx \sum_{l=1}^{N} L_j^r \left(x, x_l^{\bullet j}; h_j \right) w_l^{\bullet} y_l^{\bullet j}$$

As in the estimation algorithm, in the case of the binning technique the cross-validation error

CV_j can be approximated by:

$$CV_j \approx \sum_{r=1}^{N} w_r^{\bullet j} \left(\hat{f}_j^{-(r)} \left(x_r^{\bullet j} \right) - \frac{y_r^{\bullet j}}{w_r^{\bullet j}} \right)^2,$$

where $\hat{f}_j^{-(r)}$ is obtained without the (r) element of the binning sample.

The finer the grid of points selected, the better the binning approximations. The choice of the number of grid points is a compromise between approximation error and computational speed. In this paper, we used 30 grid points covering the range of each X_j. In practice, depending on the sample size n and the distribution of the covariates, a larger amount of grid points might be more appropriate.

References

[1] Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; **8**:183–298.

[2] Swets JA, Pickett RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press: New York, 1982.

[3] Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988; **8**: 204-215.

[4] Pepe MS. Three Approaches to Regression Analysis of Receiver Operating Characteristic Curves for Continuous Test Results. *Biometrics* 1998; **54**: 124– 135.

[5] Faraggi D. Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* 2003; **52**:179–192.

[6] Zheng Y, Heagerty PJ. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* 2004; **4**:615–632

[7] Pepe MS. An Interpretation for the ROC Curve and Inference Using GLM Procedures. *Biometrics* 2000; **56**: 352–359.

[8] Alonzo TA, Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 2002; **3**:421–432.

[9] Cai T, Pepe MS. Semiparametric Receiver Operating Characteristic Analysis to Evaluate Biomarkers for Disease. *Journal of the American Statistical Association* 2002; **97**:1099–1107.

[10] Cai T. Semi-parametric ROC regression analysis with placement values. *Biostatistics* 2004; **5**:45–60.

[11] Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.*Oxford University Press: New York, 2003.

[12] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. London: Chapman and Hall, 1990.

[13] López-de-Ullibarri I, Cao R, Cadarso-Suárez C and Lado MJ. Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics & Data Analysis* 2008; **52**:2623–2631.

[14] González-Manteiga W, Pardo Fernández JC, Van Keilegom I. ROC curves in nonparametric location-scale regression models. *Scandinavian Journal of Statistics* To appear 2010.

[15] Rodríguez-Álvarez MX, Roca-Pardiñas J, Cadarso-Suárez C. ROC curve and covariates: extending induced methodology to the non-parametric framework. *Report 09/05. Discussion Papers in Statistics and Operation Research*. Department of Statistics and OR, University of Vigo, 2009.

[16] Breiman L, Friedman JH. Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association* 1985; **80**: 580–619.

[17] Wand MP, Jones MC. *Kernel Smoothing*. Chapman & Hall: London, 1995.

[18] Fan J, Gijbels I . *Local Polynomial Modelling and Its Applications*. Chapman & Hall: CRC, 1996.

[19] Fan J, Marron JS. Fast implementation of non-parametric curve estimators. *Journal of Computational and Graphical Statistics* 1994; **3**:35–56.

[20] Pepe MS, Cai T. The analysis of Placement Values for Evaluating Discriminatory Measures. *Biometrics* , 2004; **60**: 528–535.

[21] Hanley JA, Hajian-Tilaki KO. Sampling variability of non-parametric estimates of the area under receiver operating characteristic curves: an update. *Academic Radiology* 1997; **4**: 49–58.

[22] Opsomer JD. Asymtotic properties of backfitting estimators. *Journal of multivariate analysis* 2000; **73**: 166–179.

[23] McCullagh P, Nelder JA. Generalized Linear Models. Second Edition 1989. Chapman and Hall, London.

[24] Opsomer J, Wang Y, Yang Y. Nonparametric regression with correlated errors. *Statistical Science* 2001; **16**: 134-153.

[25] Friedman J, Tibshirani R. The monotone smoothing of scatterplots. *Technometrics* 1984; **26**: 243-250.

[26] Nadaraya EA. On estimating regression. *Theory Probab. Appl.* 1964, **9**: 141–142.

[27] Watson GS. Smooth regression analysis. *Sankhya Series A*. 1964, **26**: 359–372.

[28] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.

[29] Karelis AD, St-Pierre DH, Conus F, Rabasa-Lhoret R, Poehlman ET. Metabolic and body composition factors in subgroups of obesity: What do we know?. *Journal of Clinical Endocrinology and Metabolism* 2004; **89**:2569–2575.

[30] Franzosi MG. Should we continue to use BMI as a cardiovascular risk factor. *Lancet* 2006; **268**: 624–625.

[31] Yusuf S, Hawken S, Ounpuu S, Bautista L, Franzosi MG, Commerford P, Lang CC, Rumboldt Z, Onen CL, Lisheng L, Tanomsup S, Wangai P Jr, Razak F, Sharma AM, Anand SS, on behalf

of the INTERHEART Study Investigators. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: a case-control study. *Lancet* 2005; **366**:1640–1649.

[32] Litwin SE. Which measures of obesity best predict cardiovascular risk?. *Journal of the American College of Cardiology* 2008; **52**: 616–619.

[33] Gelber RP, Gaziano JM, Orav EJ, Manson JE, Buring JE, Kurth T. Measures of Obesity and Cardiovascular Risk Among Men and Women. *Journal of the American College of Cardiology* 2008; **52**:605–615.

[34] Botana MA, Mato JA, Cadarso-Suárez C, Tomé MA, Perez-Fernandez R, Fernández-Mario A, Rego-Iraeta A, Solache I. Overweight, Obesity and Central Obesity Prevalences in the Region of Galicia in Northwest Spain. *Obesity and Metabolism* 2007, **3**: 106–115.

[35] Tomé MA, Botana MA, Cadarso-Suárez C, Rego-Iraeta A, Fernández-Mario A, Mato JA, Solache I, Pérez-Fernández R. Prevalence of metabolic syndrome in Galicia (NW Spain) on four alternative definitions and association with insulin resistance. *Journal of Endocrinological Investigation* 2008; (in press).

[36] International Diabetes Federation. The IDF consensus worldwide definition of the metabolic syndrome. *(http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf)*. Accessed 3 December 2009.

[37] Cai T, Zheng Y. Model Checking for ROC Regression Analysis. *Biometrics* 2007; **63**: 152–163.

[38] Opsomer JD, Ruppert D. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* 1997; **25**: 186–211.

[39] Opsomer JD, Kauermann G. Weighted local polynomial regression weighted additive models and local scoring. Preprint Series #00-7, Department of Statistics, Iowa State University 2000.

[40] Nielsen JP, Sperlich S. Smooth backfitting in practice. *Journal of the Royal Statistical Society Series B* 2005; **10**: 43–61.